Creativity Factor Evaluation: Towards a Standardized Survey Metric for Creativity Support

Erin A. Carroll, Celine Latulipe
Department of Software
and Information Systems
University of North Carolina at Charlotte
{e.carroll, clatulip}@uncc.edu

Richard Fung, Michael Terry
David R. Cheriton
School of Computer Science
University of Waterloo
{rhfung, mterry}@cs.uwaterloo.ca

ABSTRACT

We present a new survey metric, the Creativity Support Index (CSI) that is designed to help researchers and designers evaluate the level of creativity support provided by various systems or interfaces. We initially employed a top-down literature-based approach to develop a beta version of the Creativity Support Index (Beta CSI). We discuss our usage of the Beta CSI in three different studies and what we learned from those deployments. We also present the results from an extensive creativity vocabulary study (n=300), which revealed a set of orthogonal creativity factors. This led to the current version of the CSI presented in this paper. Initial results from these formative evaluations suggest the value of this tool in assessing and comparing creativity support tools at points in time and longitudinally.

ACM Classication Keywords: H.5.2 Information Interfaces and Presentation: User Interfaces: Evaluation/Methodology

General Terms: Measurement, Standardization

Author Keywords: creativity, creativity support tools, factorvalidation, standardized survey metrics

INTRODUCTION

Creativity support tools (CSTs) span a wide variety of domains and are often interdisciplinary in nature. However, the effectiveness of these CSTs in supporting people in creative tasks is often difficult to evaluate since creativity is not easily defined nor fully understood. There are a number of reasons why current evaluation techniques are not appropriate for measuring CSTs. Time and error metrics that are standard measures in human-computer interaction (HCI) seldom apply to creative tasks. While longer time spent on a task may normally indicate inefficiencies in a tool, spending more time on a creative task is more likely to indicate engagement with the activity, and errors in a creative endeavor are often viewed as serendipitous challenges. Another common HCI metric is productivity, but the product of creative work is often one-of-a-kind and cannot be measured against similar work to judge its merits, nor is the quantity of work produced a reasonable measure, as creative individuals are likely more concerned with quality than quantity.

Current methodologies for measuring CSTs include a variety of qualitative methods, such as observation, think-aloud studies, and interviewing. Sometimes these qualitative methods are paired with quantitative methods, such as software logging to record user behavior with a system or formal studies in which users are given creative 'tasks' to accomplish in competing interfaces. While all of these methods generate interesting results, they also have drawbacks. Qualitative methods are very rewarding but are quite expensive and time consuming. Their results afford straightforward comparative analysis, but the results from laboratory studies do not necessarily translate to the real world of creative work. Biometric measures are promising as quantitative techniques and have been used successfully in evaluating entertainment software [11], but they are expensive and can be intrusive.

Survey methods are a common tool for quantitative analysis, but there are no standardized surveys designed for measuring CSTs. Instead, HCI researchers either borrow surveys from other disciplines, which can easily alter the survey's validity, or they create custom surveys for their studies, which are not easily transferable to other studies and do not allow other researchers to replicate or validate the results. Despite the fact that survey tools are limited to self-report, we are interested in them because they are inexpensive and convenient. Standardized surveys are also beneficial in publications because the survey data will have meaning to other researchers. It is our goal to design and standardize a measurement tool that can be used in addition to other methods to help researchers in evaluating the effectiveness of CSTs.

We present a new version of our measurement tool, the Creativity Support Index (CSI), which is a revision from our beta version. The Beta CSI was based upon concepts and theories of creativity, and it was used in three different studies to test its usability. The new CSI is based on feedback from those deployments and on a principal components analysis of 300 participants' ranking of words related to creativity. Finally, we present our plans to test the new CSI and develop it into a final standardized Creativity Support Index.

C&C'09, October 26-30, 2009, Berkeley, California, USA.

CREATIVITY THEORY

An important goal of CSTs is to enhance one's creativity, since creativity research shows that an individual's creativity can be improved [1, 13]. It is our goal to measure how well a particular system or tool supports the creative activity by creating a standardized instrument. Since there are competing theories on creativity, it is important to be clear about the particular ideological foundations our instrument is built upon. Theories of creativity come from a wide range of disciplines, including humanistic (or positive) psychology, developmental psychology, and lastly, research from HCI related to creativity and CSTs. Accordingly, we also describe the primary theories of creativity that influenced the design of our research instrument and the particular features of creativity they promote. These features form the basis for the instrument we describe in the next section.

Mihaly Csikszentmihalyi studied highly creative people who are passionate about their work or hobby but are not rewarded or motivated by money or fame [3]. He attributed their passion to a concept called optimal experience or *flow*, which he broke down into nine elements:

1. There are clear goals every step of the way.
2. There is immediate feedback to one's actions.
3. There is a balance between challenges and skills.
4. Action and awareness are merged.
5. Distractions are excluded from consciousness.
6. There is no worry of failure.
7. Self-consciousness disappears.
8. The sense of time becomes distorted.
9. The activity becomes autotelic [the meaning of the activity is within itself].

This model of flow also applies to creativity, but since creativity is comprised of multiple dimensions [17], flow does not fully account for creativity. For example, we believe that creativity has a relationship to the concept of play, since creative activities are often more similar to playing than they are to working. This idea is supported by early research showing correlations between high levels of creativity and high levels of playfulness in children [13] and also by play theories. The most relevant play theory to the study of creativity is the disposition of play, as described by Rubin, Fein, and Vandenburg [16]. In their work, six factors were outlined to account for play disposition: intrinsic motivation, attention to means rather than ends, active engagement of the individual, freedom from external rules, nonliterality, and behavior dominated more by the individual than by the environment. These factors of play overlap with many of the elements in the flow model. Another play study that relates to flow comes from Read, MacFarlane, and Casey [15] in HCI. In this paper, the authors were able to document three dimensions of children's fun: expectations, engagement, and endurability. Of particular interest is their use of engagement, which was measured by observing facial expressions, and endurability, which was the willingness to continue or repeat an activity.

These models share very similar concepts, especially the idea of engagement. Engagement was found to be a dimension in both of the play studies [16, 15], and it also directly relates to the majority of the flow elements in that many of them could be outcomes of being actively engaged in a task (see elements 5, 7, and 8 in the Flow model). We also think endurability from Read et al. [15] is equally important because having a desire to repeat an activity should reflect a person's enjoyment of that task.

The HCI community has also provided important guidelines for designing CSTs. An NSF workshop on CSTs listed the following principles as critical: support exploratory search, enable collaboration, provide rich history-keeping, and design with "low thresholds, high ceilings, and wide walls" [17]. The concept of exploratory search is also a component of play and related to the concept of flow. Collaboration can be a strong component of play as well. The idea of designing with "low thresholds, high ceilings, and wide walls" can be related to the concept of freeform play in which there are no rules or where rules can be made up but are malleable.

In early developmental psychology, play and exploration are seen as completely separate behaviors [9], but we believe that they are strongly related. There is a cultural norm of equating these two behaviors, as seen in common parlance where people use the phrase, "I'm just playing around with it," to describe their activity in exploring a new device, system, or tool. It is also supported by the theory and philosophy of Montessori education, which holds that the main motivation for play in young children is the exploration of the world and that exploration is the "work of the child" [12].

The theories and concepts of creativity related to flow, play, and the creativity support guidelines formed the foundation of the Beta CSI, the first version of our measurement tool. Figure 1 synthesizes how the various definitions, concepts, and theories of creativity and play relate to the six factors in the initial version of the CSI.

THE BETA CSI

It would certainly be possible to create a very lengthy and detailed survey for measuring people engaged in a creative task, given the many concepts and theories already discussed. Such a survey would be too tedious for participants, especially for within-subject experiments. Thus, we came up with six factors through discussions, the literature research synthesis described in the previous section, and a card sorting exercise with five people in our research group. The card sorting was performed to place the terminology related to creativity theory into some general orthogonal categories. The six factors (or constructs) that resulted from the literature synthesis and the card sorting are *Exploration, Collaboration, Engagement, Effort/Reward Tradeoff, Tool Transparency,* and *Expressiveness*.

The CSI follows a similar structure (one question per factor) to the NASA Task Load Index (TLX), which is a familiar survey in the HCI community. The TLX measures work load factors, such as performance, time pressure, and frustration, and it is designed for tasks that have clearly defined objec-

Foundation of Literature in the Creativity Support Index	Exploration	Collaboration	Engagement	Effort/Reward Tradeoff	Tool Transparency	Expressiveness
Csikszentmihalyi's Elements of Flow	✓					
Clear goals every step of the way						
Immediate feedback to a person's actions		✓			✓	
Balance between challenges and skills				✓		
Action and awareness are merged			✓		✓	
Distractions are excluded from consciousness			✓			
No worry of failure	✓		✓			
Self-consciousness disappears			✓			
Sense of time becomes distorted			✓			
Activity become autotelic [meaning of activity is within itself]			✓			
Schneiderman's Design Principles for CSTs						
Support exploratory search	✓	✓				
Enable collaboration						
Provide rich-history keeping	✓				✓	
Design with low thresholds, high celings, and wide walls				✓	✓	✓
Read et al.'s Dimensions of Fun						
Expectations (reported experience better than predicted experience)			✓	✓		
Engagement			✓			
Endurability (willingness to continue/repeat activity)			✓	✓		
Rubin et al.'s 6 factors of Play as Disposition						
Instrinsic motivation			✓			
Attention to means rather than ends			✓			
Active engagement			✓			
Freedom from external rules	✓					✓
Nonliterality						✓
Behavior dominated more by person than environment		✓		✓		

Figure 1. Relationship between literature on creativity and the six factors used in the Beta CSI.

tives [7]. The TLX has been used for evaluating creativity support software [10] but is most appropriate for productivity and other related software. Creative activities must be measured differently than typical productivity work because they differ significantly. For example, artistic individuals do not necessarily set time restrictions on themselves for their creative work, so temporal demand (a TLX dimension) is likely to be a confusing question for an artist. In fact, if an artist spends a long time on an activity, this may indicate a higher level of creative engagement, rather than an ineffective tool, or interface. Similarly, mental demand (also a dimension) may be high due to the intensity of being fully engaged in a creative task, but that is reflective of the work, not necessarily an indication of a poor tool.

While the TLX does not measure the appropriate information for CST evaluation, it has many appealing features. It is easy for researchers to use and many HCI researchers are already familiar with the tool because of its standardization. Consequently, researchers are able to report TLX results without explaining all the details of the measurement tool. Participants also benefit from the TLX because they are able to fill it out quickly, which is especially important for within-subjects experiments.

Figure 2. The factor ratings page of the Beta CSI.

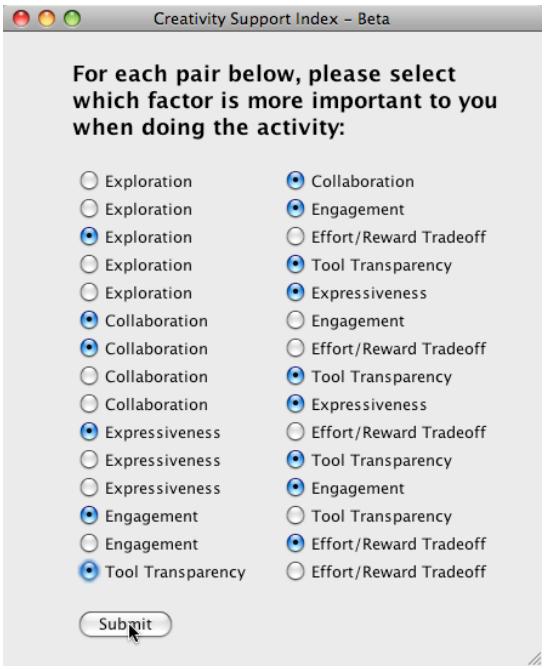

Figure 3. The pairwise factor rankings page of the Beta CSI.

The Beta CSI format and scoring is identical to the that of the TLX. The first part of the survey consists of six questions with one for each factor and has responses ranging from Highly Disagree (1) to Highly Agree (20) (See Figure 2). In the second part of the survey, each factor is compared against the other five to assess the relative importance of these factors to each participant for the activity under study (See Figure 3). For example, they will rate whether Expressiveness or Exploration was more important to them while doing the activity. Since each category must be compared with every other category, we have an additive factorial comparison of items (for 6 factors, there are 15 comparisons). We realize that the factor comparison could be tedious for a participant, as we have seen this in our experiences using the TLX; however, we believe that the factors in the CSI should be ranked independent of the actual conditions/tools being evaluated. Therefore, the factor comparison can be administered just once per participant. The overall score for the CSI is calculated by multiplying each factor rating (1-20) against the count for that factor in the pairwise rankings (0-5). This is done for each factor and then summed. Lastly, this sum is divided by 3 to arrive at an index out of 100.

BETA CSI USAGE

While we were continuing to study which factors and constructs are most appropriate for evaluating creativity, we deployed the Beta CSI in three different studies to get usability feedback on the measurement tool. Overall, the CSI survey was easy for participants to complete but two factors caused some confusion: collaboration and tool transparency.

Beta CSI Usage in Ken Burns Study

Our first usage of the Beta CSI was in an experiment aimed at determining how well two different applications support users in creating expressive slideshows from photographs. Specifically, the participants were asked to create two different slideshows using the Ken Burns Effect. This effect allows users to specify how a photograph is displayed in a slideshow by allowing them to select many regions of interest in a photograph and is a common technique in documentary filmmaking. The software animates by interpolating the viewpoint between regions of interest, allowing evocative narratives to emerge from static photographs.

In one of the experiment conditions, the participant selected Ken Burns Effect regions using two mice and two cursors to select a rectangular area of interest (similar to a cropping tool), and in the other condition, they specified the Ken Burns Effect regions using panning and zooming (similar to the interface in Apple's iPhoto software). The participants could select as many Ken Burns Effect regions as they wanted, and the application then produced animated slideshows by interpolating the viewpoint between each Ken Burns Effect region. After using each technique to create a slideshow, they were given both the TLX and the CSI with survey order counterbalanced across participants.

This study was a within-subjects setup with 32 participants recruited from the psychology department's participant pool, where they were given research credit for participating. In this study, participants filled in the CSI survey electronically. The average CSI score was 72.94 (out of 100) in the two mice condition and 62.81 in the panning and zooming condition, but these CSI scores were not significantly different (t[30]=-1.65, p=0.10). There was a high standard deviation between the pan-zoom condition (*SD*=26.3) and the dual-cursor condition (*SD=19.2*). This shows that while most participants' responses generated a higher CSI score for the dual-mouse condition than for the pan-zoom condition, the response levels varied significantly across individuals. The use of a 20-point rating scale allowed for significant variation in the ratings for each individual category. However, we do not put a lot of emphasis on these numbers because the CSI was in its beta version. We were most interested in the subjective responses of the participants when taking the survey, which is discussed next.

CSI Feedback

There were two questions that seemed to cause confusion with participants: the question on collaboration and the question on tool transparency. The collaboration question was given the factor heading "Collaboration," followed by the statement, "I was able to work together with others easily while doing this activity." We chose to include collaboration in the CSI, even though not all CSTs support collaboration. Our reasoning is that if a task does not require collaboration, participants will not choose collaboration in the pairwise rankings, and thus, the actual rating they give to collaboration is not important. Designing the survey this way allows us to keep the collaboration factor in the survey so that it can be part of the equation when it is relevant. When collaboration is not relevant, it does not effect the equation because of the pairwise rankings. The goal that led to this design decision was the desire to develop one CSI survey, rather than separate versions for collaborative vs. non-collaborative tools. Measuring collaboration would also be important in evaluating CSTs where only one tool may have collaboration support.

Since no collaboration was involved in the slideshow creation task, we expected that participants would either choose a neutral ranking (10 out of 20) or a negative ranking (1 out of 20) for this factor. We also expected that they would not select collaboration in the pairwise rankings very often, as most people would be accustomed to software designed for individual use and would not consider collaboration to be important. As explained above, we anticipated that it did not matter what rating participants gave for the collaboration factor, as we thought they would not rank collaboration as particularly important to the task. Therefore, we expected the ratings (when multiplied by low or zero rankings) would be largely irrelevant to the overall calculation.

During the Ken Burns study, a few participants verbally asked if they should ignore the Collaboration question or if it was not applicable, which indicates that the question did cause some confusion. Our estimation was partly true: nine participants selected the lowest rating and eight selected the middle value. Three selected the highest rating and the other eleven participants were scattered across other values. As

for the pairwise rankings, the average count for collaboration (which could range from 0 to 5) was 0.9, which supports our idea that participants would not rank collaboration as important and that their ratings of collaboration would therefore not strongly affect the CSI. It is possible that the collaboration issues stemmed from a misunderstanding of the question. The phrasing of the question did not focus on the collaborative affordances of the tool but rather on the collaborative nature of the activity. Even though participants did not work with others on the task, they may have interpreted it as asking how well they could imagine themselves collaborating with others on the activity.

The other question that seemed to cause some confusion was the "Tool Transparency" factor, which used the statement, "While I was doing the activity, the tool/interface/system 'disappeared,' and I was able to concentrate on the activity." While the data for Tool Transparency did not show any indications there was a problem, several participants did ask for clarification on the meaning of the question, indicating that this question was also a source of confusion. It appeared that participants were taking this question literally and were trying to recall if there were dialog boxes or other controls that were literally transparent.

Survey Comparison

What is especially interesting about this particular study is that the users also filled out the TLX survey, so we were able to do some cross-survey comparisons. The TLX overall scores were almost identical for the two conditions, and thus were not informative. At the end of this study, we asked meta-questions about the surveys completed by the participants, in an attempt to determine end-user perceptions of the two survey tools. We asked participants which of the two surveys was the most appropriate survey for the slideshow creation task they had just performed. Twenty one participants selected the CSI as the most appropriate survey for these tasks, while only 10 selected the TLX as the most appropriate survey. One participant did not complete this final set of questions.

During this experiment, we also asked participants which survey, if any, they found most confusing. The TLX was identified by 14 participants as most confusing, four participants said both surveys were confusing, and 13 participants said neither were confusing. None of the participants identified the CSI as most confusing. These results are a positive indication that in the task context, the CSI survey makes more sense to participants than the TLX.

Beta CSI Usage in Color Exploration Study

The Beta CSI was also used in a smaller think-aloud study for a bimanual color exploration tool. This was not a comparative study but was aimed at getting feedback about a new technique from a specialized set of users. Here, rather than users from a general population, the eight participants were all digital artists, architects, or designers. The study was run as a one-hour session where they were video-recorded and asked to think aloud about their experience using the application and the color exploration tool. At the end of the hour, they were asked to complete the Beta CSI survey by filling out the survey on paper. As there was no control condition, there are no comparisons to be made. In this instance, we were interested in looking at the aggregate categorical data.

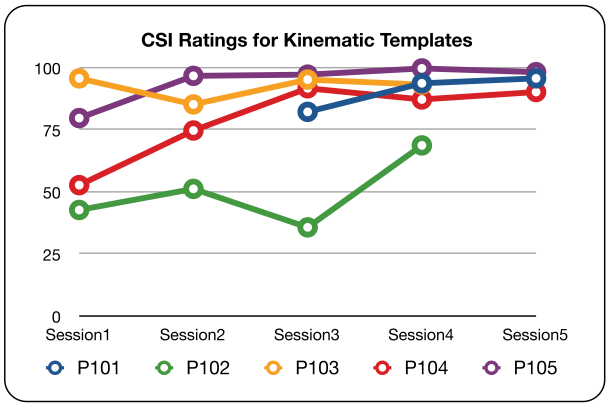

Figure 4. Beta CSI scores for the five participants and five sessions in the Kinematic Templates study. Note that not all participants attended all sessions, but there is a general trend towards higher CSI scores with repeated exposure to the kinematic templates application.

As with the Ken Burns study, we noticed that participants were confused by the collaboration question and by the tool transparency question. Two of the eight participants wrote "N/A" beside the Collaboration question. One participant also wrote a note beside the Tool Transparency question that said "Yes, it disappeared, but it would have been easier if it stayed." This participant had mentioned during the study that she wanted to pin open the dual-cursor color exploration dialog box. This demonstrates once again that the transparency question is being interpreted too literally.

Beta CSI Usage in Kinematic Templates Study

The third usage of the Beta CSI was in a study of a drawing program that makes use of varying control-display gain ratios to allow a variety of interesting kinematic drawing effects [4]. An evaluation was performed to determine where and how templates could be useful. Artistically-inclined volunteers were invited to produce visual compositions with kinematic templates. Each individual participated in four or five sessions, each lasting approximately one hour, over a course of three to twelve weeks. Thus, this was a longitudinal study in which the CSI survey was administered on paper after each session. The pairwise factor ranking was done only once, after the first session. In this study, the Beta CSI was altered after the first session to remove the collaboration factor (both from the ratings and the pairwise rankings) based on participant confusion with that question. Removing collaboration as a factor allowed us to see how the survey worked without it.

This usage of the CSI was interesting because it showed how the survey could be deployed in a more longitudinal study, which may be essential for the evaluation of new creativity support tools that are complex. In general, the Beta CSI scores increased with each session in this study, indicating

an improved experience with using kinematic templates over time, as shown in Figure 4. The participants seemed to become more comfortable with the software after learning the interface and as bugs were fixed. Additionally, in the latter sessions, participants were allowed to draw anything they wanted, which increased their engagement level, as reflected in increasing Engagement ratings over the five sessions.

One of the participants who produced impressive drawings and clearly had a strong interest in art, asked about the Exploration factor because of a phrase in the statements wording: "...without a lot of tedious, repetitive interaction." The participant pointed out, "I kind of like tedious, repetitive interactions... it's just the way I draw." This participant was observed to use the same action or template repeatedly to draw a particular feature but was not exploring different alternatives. This feedback indicates that rephrasing the exploration question may be warranted. The other issue we were interested in was collaboration. We removed the collaboration factor from the ratings given that the Kinematic Templates interface did not directly support collaborative interaction. However, by not including collaboration in the pairwise rankings, we were unable to determine whether collaboration support was important to the artists.

Factor Ranking Graphs

Deploying the Beta CSI in three unique situations allowed us to obtain valuable feedback. One aspect of the results that could be quite useful is the counts gained through the pairwise factor rankings. As mentioned previously, the participants were asked to select from each pair of factors which is most important to them when engaging in the activity under study. As there were six factors in the Beta CSI, there were 15 comparisons to be made and therefore 15 points to be awarded across the six factors. By dividing the average counts for each factor by 15, we get a percentage weighting, which indicated the relative importance of each creativity factor to the activity for the users under study. Figure 5 shows two pie charts illustrating these results for the slideshow activity (general population) and the color vector drawing activity (artists/architects/designers). These charts can also illustrate the possibility of helping designers understand which aspects of creativity support are particularly important across different creative activities or across different types of users. For example, factor comparisons across a population could show what factors are more important to novices vs. experts in a creative activity.

CREATIVITY RATINGS

After deploying the Beta CSI in three studies, we were able to identify a number of issues that needed to be addressed. First, our participants may have different subjective views of creativity than the terminologies and concepts from creativity research, and second, that our card sorting process may not be sufficient to form the survey factors of creativity.

In order to form the new constructs, we used a similar process that was used by Hart and Staveland in developing the TLX [6]. In their early research, the authors presented participants with words from research on workload and asked

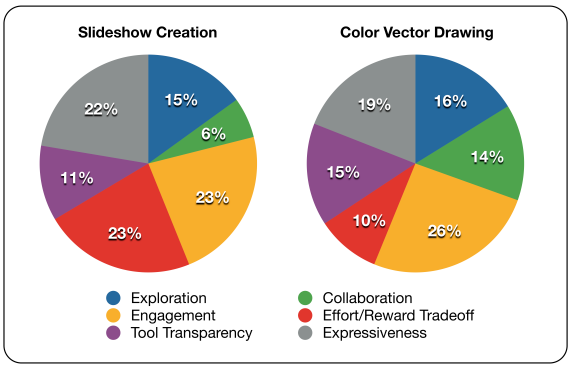

Figure 5. Relative importance of each creativity support factor for a general population creating a slideshow and for artists/architects/designers creating vector graphic drawings.

them to rate to what extent the word was related to workload. From here, they analyzed the data with a principle components analysis (PCA), which is used for reducing data and forming constructs. Specifically, it is used to extract all the components that account for the majority of the variance in a data set. After this extraction, the significant loadings (correlations) between the variables in each component are identified and can be renamed to represent a construct.

In our study, we recruited 300 people with a wide distribution in age, gender, and ethnicity using Amazon's Mechanical Turk [2]. We presented them with the 19 words displayed in Table 6. For each word they were asked to indicate whether it was *extremely important*, *important*, *somewhat important*, or *not at all important* to the creative process.

The Mechanical Turk raters were also required to complete the Originality/Creativity scale from the International Personality Item Pool (IPIP). This consisted of seven questions where they rate their agreement using the categories: Very Inaccurate, Moderately Inaccurate, Neither Inaccurate nor Accurate, Moderately Accurate, or Very Accurate [5] (See Table 1). Scores for this survey are obtained by adding the values from all questions together, but the scale must be reversed for negative questions. While the IPIP is obviously limited to self-report, it provided us with a standardized, objective (and obviously estimated) measure of each participant's creativity level. The idea was to group the participants according to their creativity and then determine whether that has any correlation with the ratings of the 19 words.

RESULTS FROM CREATIVITY RATINGS

The ratings from the 19 words representing creativity showed that there were widespread definitions for creativity, as seen in Figure 6. The only word that scored low was collaboration, which was rated as being essential to the creative process by only 35.6% of the raters. We expect that we received these results because the average person may not include collaboration in the majority of their creative endeavors, thus they may not have seen collaboration as essential to creativity. There also seems to be a cultural stereotype of the creative genius working alone in their studio or

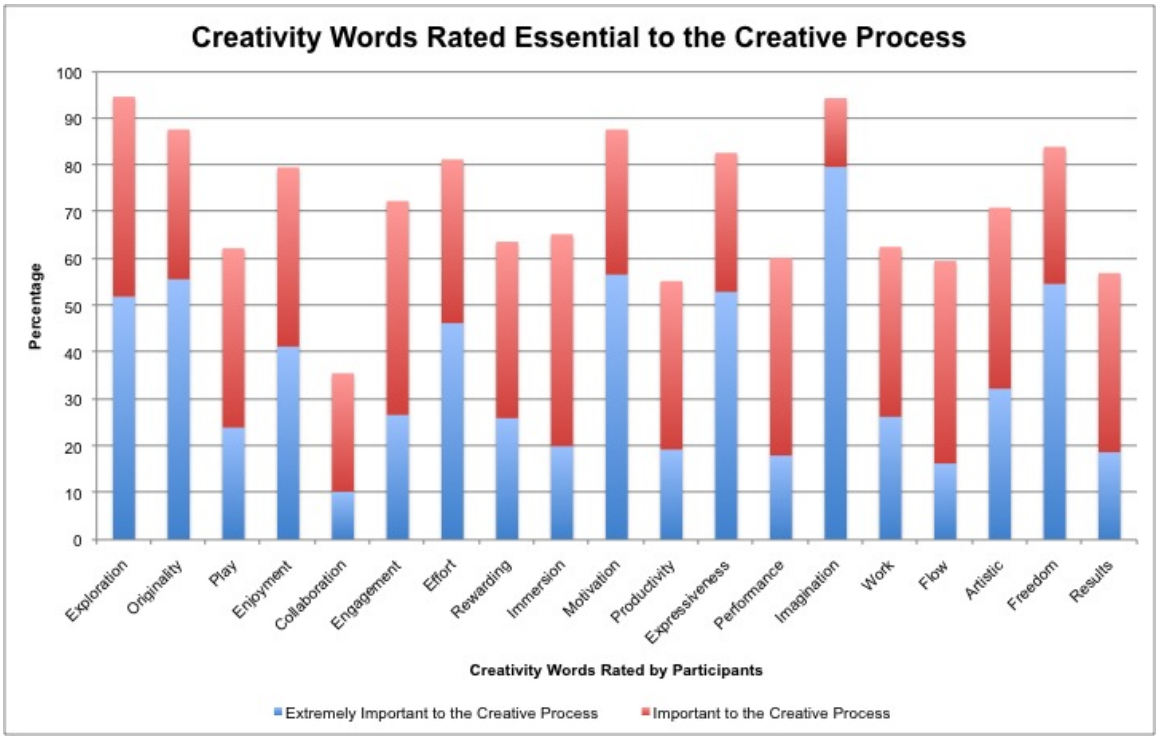

Figure 6. These percentages correspond to the rankings of the 19 creativity words that were rated as being essential to the creative process.

Originality/Creativity Questions from the International Personality Item Pool (IPIP)
1. I am able to come up with new and different ideas.
2. I have no special urge to do something original.
3. I come up with new ways to do things.
4. I have an imagination that stretches beyond that of my friends.
5. I don't pride myself on being original.
6. I am not considered to have new and different ideas.
7. I am an original thinker.

Table 1. These questions from the International Personality Item Pool (IPIP) were answered by the creativity word raters, where they selected their level of agreement with each question.

Chi-Square Results of Significant Factors					
Factor	df	n	X^2	p	r
Play	66	289	105.10	.00	0.15, p=0.01
Enjoyment	66	292	86.34	.05	0.16, p= 0.01
Expressiveness	66	291	109.90	.00	0.19, p=0.00
Flow	66	293	89.18	.03	0.09, p=0.14
Freedom	66	292	100.70	.00	0.19, p=0.00
Results	66	294	93.59	.01	-0.10, p=0.07

Table 2. These factors were shown to have significant score differences based on IPIP scores. All of the factors, except flow and results, also had significant correlations.

lab on an invention, which persists despite much research that shows many creative inventions were a result of joint effort [8]. It is also possible that since Support Collaboration is a design principle in CST research, we should have asked them whether supporting collaboration was essential to creativity, not just collaboration itself [17].

The IPIP scores were also examined to see if an individual's creativity affected which words they considered essential to creativity using two statistical tests. We used both Pearson's correlation and Pearson's chi-square test. The chi-square test is used for testing the null hypothesis in a frequency distribution and was selected because we had categorical variables. The results of our chi-square test found that IPIP scores significantly affected the distribution of ratings for six of the creativity words. Of these words, all but two had a significant, positive correlation with the IPIP score. Specifically, people with higher IPIP scores rated play, enjoyment, expressiveness, and freedom as significantly more important to the creative process (See Table 2).

Our word ratings results revealed no conclusive definition of creativity. Therefore, our goal was to reduce the data into smaller dimensions by finding which words were correlated, as was done in the TLX [7]. We first used a statistical test called the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and found heavy correlations across the creativity words (*KMO*=0.76). In a KMO, anything above 0.60 would indicate that correlations in the data may exist and serves as a good pre-test for running a PCA. In other words, if the KMO revealed no correlations or very low ones, then the PCA would more than likely produce poor results.

Since the KMO indicated correlated variables and since the

Rotated Components Matrix of the Principle Components Analysis						
	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
Exploration	0.024	0.009	0.288	**0.708	-0.177	0.044
Originality	0.135	0.306	0.206	0.045	0.119	-0.703
Play	-0.077	*0.414	-0.081	*0.541	0.217	-0.015
Enjoyment	0.111	*0.423	0.241	0.093	0.142	*0.597
Collaboration	*0.507	-0.087	-0.159	*0.413	-0.005	0.238
Engagement	0.153	0.026	-0.163	*0.593	0.338	-0.079
Effort	*0.590	-0.104	*0.409	0.060	0.186	0.010
Rewarding	*0.459	0.098	0.172	-0.049	0.296	0.303
Immersion	0.085	0.040	0.152	0.079	**0.859	-0.040
Motivation	0.250	0.016	**0.748	-0.090	0.266	0.039
Productivity	**0.639	0.192	0.130	-0.148	0.155	0.028
Expressiveness	-0.029	**0.729	0.278	0.027	-0.068	-0.058
Performance	**0.763	0.188	-0.019	0.136	-0.043	0.017
Imagination	-0.121	0.320	**0.718	0.107	-0.133	-0.078
Work	**0.682	-0.176	0.027	0.173	0.160	-0.020
Flow	0.292	*0.410	-0.147	0.122	*0.438	0.144
Artistic	0.180	**0.740	-0.073	-0.082	-0.017	0.066
Freedom	-0.012	*0.579	0.135	0.177	0.190	-0.118
Results	**0.730	0.092	-0.062	-0.060	-0.059	-0.159

Table 3. The principle components analysis (PCA) extracted six components of correlated factors from the 19 creativity words. ** refers to heavy loadings (> 0.60), * refers to low loadings ($0.40 - 0.60$)

Words	Construct	Beta CSI
collaboration, effort, **work, productivity performance,** rewarding, **results**	Results Worth Effort	Effort/Reward Tradeoff
play, enjoyment, flow, **expressiveness,** freedom, **artistic**	Expressiveness	Expressiveness
effort, **motivation, imagination**	*n/a*	*n/a*
exploration, play, engagement, collaboration	Exploration	Exploration
flow, **immersion**	Immersion	Tool Transparency
enjoyment	Enjoyment	Engagement

Table 4. Extracted components from the PCA for the 19 creativity words. The construct column shows how we have named the components to reflect the correlated variables, and the Beta CSI column shows how they map to factors in the Beta CSI. Bold indicates heavy loadings.

PCA was also a primary statistical test used in the TLX [6], we ran a PCA. This revealed that the raters' view of creativity reduces to six components accounting for 58.20% of the data's variance. It is a coincidence that the PCA extracted the same number of factors used in the Beta CSI. In a standard PCA, components that are extracted must have an eigenvalue of at least 1, so this was the procedure that we followed. Results of the PCA are available in Table 3, which shows the loadings across variables in each component.

The six extracted components from the PCA have been renamed to represent the constructs that account for the correlated variables (see Table 4). These PCA results will be used in the new CSI with two minor modifications. First, the third component, which has 'motivation' and 'imagination' as the heavy-loading terms, will not be included in the CSI. We have decided to exclude these words because they are intrinsic to personality. If researchers are interested in measuring these individual characteristics, personality metrics, such as IPIP scales [5], are more appropriate.

Our second modification involves collaboration. As seen in Table 4, collaboration appeared in two components but was not heavy loaded. Thus, there is no component from the PCA that particularly maps to the Collaboration factor that we included in the Beta CSI. One option would be to remove this factor, but we do not believe that is a good idea. There are strong reasons to believe that collaboration support is an important aspect in supporting creativity, so we are keeping Collaboration as a separate factor in the new CSI. Thus, our final components are: Results Worth Effort, Expressiveness, Exploration, Immersion, Collaboration, and Enjoyment.

THE NEW CSI

We have revised the CSI to include updated factors and statements based on the results of the Beta CSI usage and the PCA results. These changes are expected to address construct validity and reduce confusion for future participants. We present here the new version of the tool (See Figures 7 and 8), along with explanations of the changes in the tool.

Results Worth Effort: *What I was able to produce was worth the effort I had to exert to produce it.* Words associated with this generated the most prominent component in the PCA. We have changed this label to 'Results Worth Effort' to make it more straighforward. By replacing the term 'reward' with 'results' we hope to be less ambiguous,

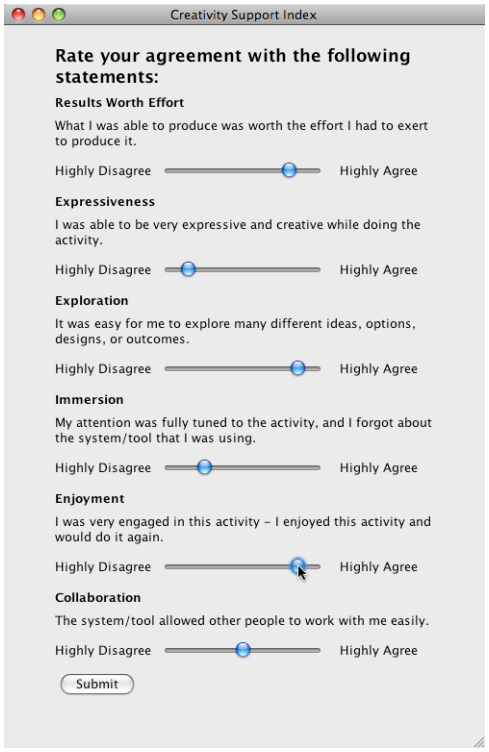

Figure 7. The factor ratings page for the CSI. The sliders generate ratings from 0 to 10.

as some people might not consider what they produce in a creative activity as a 'reward.'

Expressiveness: *I was able to be very expressive and creative while doing the activity.* Expressiveness was a clear component of the PCA and will remain in the CSI. While 'artistic' was a high-loading component in the same component, we chose to not replace the term 'creative' with the term 'artistic' since 'artistic' seems too narrowly focused on the visual arts domain.

Exploration: *It was easy for me to explore many different ideas, options, designs, or outcomes.* The Exploration factor was evident in the PCA, so it will remain in the CSI.

Immersion: *My attention was fully tuned to the activity, and I forgot about the system/tool I was using.* Immersion is exactly what we meant by our 'Tool Transparency' factor, but immersion is much less ambiguous.

Enjoyment: *I was very engaged in this activity - I enjoyed this activity and would do it again.* When we originally formed the factors for the Beta CSI, we had separate factors for 'enjoyment' and 'engagement' but decided that these were not orthogonal and thus used engagement instead. However, the PCA results indicate that people may be more comfortable associating the term 'enjoyment' with creativity, so we have replaced 'engagement' and modified the survey statement.

Collaboration: *The system/tool allowed other people to work with me easily.* Although collaboration did not come

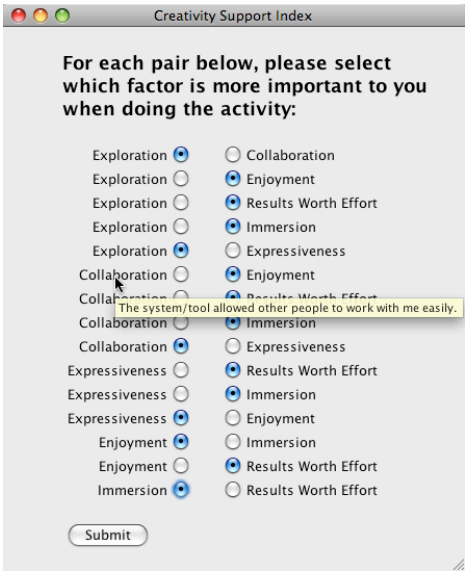

Figure 8. The pairwise factor rankings page for the CSI. Notice that in the electronic version of the survey, the statements describing each factor pop up when the user hovers the cursor over each factor.

up as a separate component in the PCA, we believe that it is a very important aspect of creativity support that must be measured individually. We have phrased the question to be less ambiguous, focusing on the collaborative aspect of the system or tool.

The factors outlined here will first be rated and then ranked through pairwise comparisons, as in the Beta CSI. While the beta version used a rating scale of 1-20, the new version will use 0-10. This change was made because test-retest validity is higher when there are more points on the rating scale (greater than 7) but validity decreases when scales go above 10 points [14]. In order to arrive at a Creativity Support Index that is between 0 and 100, the summed values must be divided by 1.5, as shown in the equation below.

$$\begin{aligned} CSI = \quad & (Exploration * ExplorationCount \quad + \\ & Expressiveness * ExpressivenessCount \quad + \\ & Immersion * ImmersionCount \quad + \\ & EffortResults * EffortResultsCount \quad + \\ & Enjoyment * EnjoymentCount \quad + \\ & Collaboration * CollaborationCount)/1.5 \end{aligned}$$

CSI SUMMARY

The CSI has more validity through its evolution from the Beta CSI which was tested in three different types of studies. Those results show that this measure can be useful for looking at point-in-time interface comparisons as well as longitudinal comparisons. We have also shown that the CSI is useful for determining the relative importance of various factors in a particular creative activity, regardless of the CST being evaluated. The factors in the CSI have been validated through the 300 participant word rating study and the prin-

cipal components analysis. Thus, the CSI is ready to deploy on a larger scale. We aim to test the CSI tool in a variety of situations, and we invite other researchers to use the tool, as well. While we anticipate that there may be minor tweaks made to the instrument, we believe that this version of the CSI is close enough to our final release to be generally useful. We will do more usability testing to ensure that the factors and statements are comprehensible to participants and we also plan to do reliability testing through test-retest studies. The tool clearly needs to be evaluated in situations where there is participant collaboration. After further usability and reliability testing and more deployments, we expect that the CSI will be adopted as a standard metric by designers and researchers working on creativity support tools.

CONCLUSIONS AND FUTURE WORK

We have presented a new measurement tool for evaluating creativity support: the Creativity Support Index (CSI). The CSI encompasses six orthogonal factors related to creativity support: *exploration, expressiveness, enjoyment, immersion, collaboration* and *results worth effort*. The survey metric generates an index between 0 and 100 of the creativity support afforded by a system, tool, or interface.

The CSI evolved from a pilot version, the Beta CSI, which was grounded in concepts and theories of creativity and was deployed in three different studies to gather feedback on its usability as a research metric. Participants completing the Beta CSI had issues with some of the factor statements, which have been addressed in the new version. We also presented results from a separate study of 300 participants' creativity word ratings. The PCA of the creativity word ratings generated six orthogonal components that were incorporated into the new CSI with slight modifications.

While the CSI is a self-report tool and carries with it all of the issues that are associated with self-report measures, we believe it will be an important and useful metric that can complement other evaluation methods. We invite other researchers to try this tool [1], and we welcome their feedback.

The next step in this research is to further validate the CSI. We will deploy the CSI in some studies to ensure that we have addressed the usability issues identified with the Beta CSI. We then plan to focus on further formal validation of the CSI through test-retest reliability studies, in conditions with and without collaboration.

REFERENCES

1. T. M. Amabile. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357–376, 1983.
2. Amazon. Amazon Mechanical Turk, 2009.
3. M. Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perrennial, 1997.
4. R. Fung, E. Lank, M. Terry, and C. Latulipe. Kinematic templates: End-user tools for content-relative cursor manipulations. In *Proceedings of ACM UIST 2008*, 2008.
5. L. Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. D. Fruyt, and F. Ostendorf, editors, *Personality Psychology in Europe*, volume 7, pages 7–28. Tilburg University Press, Tillburg, The Netherlands, 1999.
6. S. G. Hart, M. E. Childress, and J. R. Hauser. Individual definitions of the term "workload". In *Eighth Symposium on Psychology in the Department of Defense*, pages 478–485, 1982.
7. S. G. Hart and L. Staveland. Development of the NASA TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, pages 239–250, 1988.
8. M. J. Howe. *Genius Explained*, chapter Inventing and discovering, pages 176–187. Cambridge University Press, United Kingdom, 2000.
9. C. Hutt. Exploration and play in children. In *Play, exploration, and territory in mammals*, volume 18, pages 61–68, London, 1966.
10. C. Latulipe, I. Bell, C. L. Clarke, and C. S. Kaplan. symtone: Two-handed manipulation of tone reproduction curves. In *GI 2006 Proceedings*, 2006.
11. R. Mandryk. Objectively evaluating entertainment technology. In *CHI '04*, pages 1057–1058. ACM Press, 2003.
12. M. Montessori. *The Absorbent Mind*. Henry Holt and Company, 1995.
13. R. S. Nickerson. Enhancing creativity. In R. J. Sternberg, editor, *Handbook of Creativity*, pages 392–430. Cambridge University Press, New York, NY, 1999.
14. C. C. Preston and A. M. Colman. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1–15, 2000.
15. J. Read, S. MacFarlane, and C. Casey. Endurability, engagement, and expectations: Measuring children's fun. In *IDC '02*, volume 53-64. Shake Publishing, 2002.
16. K. Rubin, G. Fein, and B. Vandenburg. Play. In P. Mussen and E. Hetherington, editors, *Handbook of Child Psychology*, pages 693–774. John Wiley and Sons, Inc., Wiley, New York, 1983.
17. B. Schneiderman, G. Fischer, M. Czcerwinski, B. Myers, and M. Resnick. Creativity support tools: Report from a US National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction*, 20(2):61–67, 2006.

[1] Available at `http://www.celinelatulipe.com/Home/Creativity_Support_Tools_Survey.html`